Network Analysis of Wikipedia Categorization and the Influence of Article Quality

Network Data Analysis - Final Project

Bhavik Naik
Faculty of Business & I.T
Ontario Tech University
Oshawa, Ontario, Canada
bhavik.naik@ontariotechu.net

Keng Wai Van
Faculty of Business & I.T
Ontario Tech University
Oshawa, Ontario, Canada
keng.van@ontariotechu.net

ABSTRACT

In this final report, we present a background understanding of how Wikipedia, its articles, editors, and collaborations function through a series of ten research papers. We then implement our own analysis of Wikipedia's link network in relation to their categorization community. Finally, we discuss what quality means on Wikipedia and our findings of why Wikipedia should be classified as a trustworthy source

KEYWORDS

Wikipedia, research, quality, community, network, writers

1. Introduction

Wikipedia is a large, free and online source of information. It uses a coauthorship method of article creation, and as such creates social networks. As Wikipedia continues to grow, and well surpasses its predecessor, the Encyclopedia, it becomes a valuable source of information for social network analysis. This is evident in the amount of research that has been put into using Wikipedia to analyze social networks ranging from being about Wikipedia editors' relationship to articles, to pharmaceutical companies' relationships to diseases in regards to their hyperlinks in the articles. Thus we hope to further this study into Wikipedia with our own analysis into its categorization. We then discuss the results of the analysis in aspects such as communities and centrality.

2 Related Papers

In this section, we will go through ten (10) related papers that focus on the Wikipedia networks and how various connections are formed on the online encyclopedia.

2.1 Measuring author contributions to the Wikipedia

The first paper that we will be covering is written by B. Thomas Adler, Luca de Alfaro, Ian Pye, and Vishwanath Raman, all part of UC Santa Cruz. This paper was part of the Proceedings of the 4th International Symposium on Wikis in September 2008. The goal of this paper was to understand the quality measurement in Wikipedia articles, and to figure out how to create a way to determine if Wikipedia authors are creating articles that are improving the quality of the site. The authors introduce a metric called "total edit longevity," which they claim can improve the ability to calculate the quality of the work.

In the introduction and related work sections, the Adler et al. [2008] go over past papers on this topic and discuss why previous calculations did not take into account writers who spent time polishing the articles, and those that have manipulated their numbers. They also state that previous metrics such as total text created and the number of edits made by that author does not paint a clear picture of the overall quality of that author. In Wikipedia's case, this metric of quality is a massive deal for editors as they can claim awards, promotions, and even revenue from the site. Writers that dedicate their lives to writing articles for the free encyclopedia should be awarded for their work. Despite this, many writers do abuse the system by implementing one line edits and quickly retracting them, leading to their edit count increasing, but their quality staying the same. Previous works also looked at the number of citations a page has, which could indirectly measure how much work that author has done to write that article. Later in the introduction, the researchers talk about how their favorite metric is "edit longevity" and "text longevity". Edit longevity deals with how many edits or changes that author has made and how long that edit has been left. This metric also includes the edit size, which is how many words/characters were changed. Text longevity is broken down into how much text the author has put on Wikipedia, and how long that text remains after numerous edits. Finally, the other metric that the authors introduced was, "Text longevity with penalty." This metric is for penalizing editors that try to "game" the system by editing and then retracting edits. It is important to note these authors as we can filter them out later on.

In the definitions section, Adler et al. [2008] go over the various formulas that they were using to determine the metrics. For the quantity measures, determining the amount of text is fairly easy, however for the edit distance, the authors focused on the number of words that the author either added, subtracted, or moved from the article. For the quality aspect, the first metric that needs to be calculated is the quality of edits that were performed by that author in that change. This is indicated by which edits remain unchanged after numerous revisions of the article. They then proceed to define the average edit quality of the author, which takes in all of the edits made by that author. The final two metrics that they define are the two different types of text quality measures. The first one focuses on text decay, or how much of the text remains from that author in the final copy. If no text is changed, then the quality of that text is 1, and it will decrease after each change to that text. For the second method, the equation merely calculates how much of the text remains unchanged after 10 revisions.

In the contribution measures section, the authors break down the metrics that they want to focus on to calculate the quality of that author. They wanted to take in both the quantity of work, the number of edits/words added, and also the quality of text, using the definitions that we outlined above. Adler et al. [2008] picked all revisions of articles that were published before October, 2006. The values that they wanted to focus on were, number of edits, number of words that were added, the edit distance (as outlined the above), the text longevity or text decay, the edit longevity (how many words/edits remain after 10 revisions), and the text longevity with penalty.

Onto the analysis and results, the authors outline all of the metrics in the previous paragraph and try to make sense of the metrics that they created. They found that their “Edit Longevity” metric was the best metric to determine the involvement of the editors on Wikipedia. This metric allowed Adler et al. [2008] to determine which writers deserve promotions and revenue as it takes into consideration the edit distance and the longevity of the edit. The edit distance paired with the edit longevity allowed them to filter out the best Wikipedia contributors. They also concluded that the “Text Longevity With Penalty” metric is accurate enough to find the writers that are trying to game the system. Interestingly, the highest author for the penalty metric was a bot. That being said, Adler et al. [2008] stresses that not all bots are developed for the purpose of harming articles, as there are many that help clean up text and help with formatting.

Overall, this paper succeeds in trying to research a metric that can determine the quality of writers. By creating the “Edit Longevity” measure, the authors have developed a new way to understand how Wikipedia is run and how editors on the platform are motivated to write for the free encyclopedia.

2.2 Measuring Wikipedia

The second paper that we will look at is by Jakob Voß, a part of the Humboldt-University of Berlin. This paper was published in the International Conference of the International Society for Scientometrics and Informetrics in 2005. The paper goes into what Wikipedia is, the various functions it possesses, and then talks about how it resembles a scale free network.

In the introduction and history sections of this paper, the author goes over the history of Wikipedia and how it has over 1.5 million articles. Of course, this number has significantly grown since 2005 and it is now over 55 million [1]. The author points out that as Wikipedia is fully accessible to everyone, it is easy to gather information about the site. This makes the process of understanding it far easier, and it can lead to more thoughtful findings. Also, every edit can be traced back to a user, and each revision is fully accessible by those that are reading the articles. The author also dives into the history of Wikipedia, which project it comes from, and the various languages it includes for users.

The research section overviews past works that try to understand Wikipedia and the networks surrounding it. The author overviews Viégas’, Wattenberg and Kushal (2003) paper on how they were able to visualize the edit networks within Wikipedia. They also cover the social network on collaborative writing, and how it relates to Wikipedia’s growing collection of writers.

The paper then shifts into talking about Wikipedia’s structure, how it is growing, the numbers that try to illustrate how large the network is, and components that are fundamental to Wikipedia. Starting with growth, the author points out that Wikipedia is growing at an exponential rate. The number of writers and articles have been increasing significantly and freelance article writing jobs are persuading more people to contribute. Voß [2005] outlined six metrics to track this growth and they included, database size (bytes), number of words, internal links, articles, active authors, and very active authors. These metrics allowed the author to lay out 3 phases of growth for Wikipedia. Of course, as this was written in 2005, the number of phases have increased and there have been many leaps forward for the platform. The three phases of growth that the author noticed was a linear growth initially with 10 active writers, then an exponential growth phase, and finally back down to a linear growth. This 3 phase growth is apparent in all languages that Wikipedia was formulated in.

In the article section, Voß [2005] talks about the various features of an article on Wikipedia, such as the talk feature, that allows discussions to occur on the site. It also explores the other languages that were available at that time and covers the different values of metrics describing how engaged the audience are when it comes to Wikipedia in their language. The article subsection concludes by stating that article sizes were lognormally distributed and as Wikipedia kept increasing in size, the longer the articles became.

The author section then focuses on the contributors to the free online encyclopedia. The number of unique editors on the site followed a power law distribution.

Voß [2005] also states that articles that are edited by multiple writers usually contain 4-5 editors and also follows a power law distribution. This power law distribution is also apparent when comparing the number of articles one writer will work on. An editor might create an account just to edit one article, and Voß [2005] theorizes that with a website like Wikipedia, the chance of this happening is far greater than those editors that have accounts to edit articles every day. The final discussion about authors focused on trying to determine the number of unique authors and the demographics of these writers. The average age of a writer was 31 on the German Wikipedia site.

The next few sections focused on the edit, linking, and content structure. The paper then concludes with a discussion on quality. Back when this paper came out, there was an average of 16 edits per minute on the English site and Voß [2005] also explained on a high level on how an edit can be measured. It involves calculating the similarities between revisions and using compression algorithms to calculate how much of content was changed. The link structure in Wikipedia follows a power law distribution network as pages that have less than 5 outgoing links tend to be deleted. Broken links are also something that has to be taken into consideration as many articles reference something that does not have an article of its own. As this paper was written in 2005, it talks about navigation and better markup, all items that exist in today's version of Wikipedia. Finally, for the quality section, the author explores previous work in the field and how the edit history is too time consuming to decipher. Previous work looked at the number of edits, number of writers, and other metrics to determine quality. Vandalism is another area which the author pinpointed that needed to be improved if Wikipedia was to be successful.

This paper was the first few that looked deeper into Wikipedia and the network that was created by a global community of contributors. It explored all structures and objects that make Wikipedia what we know of it to be nowadays.

2.3 Size matters: word count as a measure of quality on wikipedia

The third paper that we will overview in this report is written by Joshua E. Blumenstock, a part of the University of California at Berkeley. It was written in 2008 and published in the Proceedings of the 17th international conference on World Wide Web. The goal of this paper was to prove that word count is a great metric to determine the quality of the article on Wikipedia. They claim that it outperforms other, more complex calculations.

The paper first begins with an introduction about the quality of Wikipedia articles and how articles that get featured on the Wikipedia homepage have to go through a robust review before it is classified as an excellent article. Blumenstock [2008] wanted to simplify this process and they thought of a simple metric that can determine the quality of an article, word count. They claim that previous works were qualitative, complex, and did not paint a clear picture of the actual quality of the work.

Moving to the methods section, Blumenstock [2008] mentions that by using word count, it is easy to measure, does not involve gathering information that may be difficult to find, and that other methods are hard for the average reader to understand. To test the word length metric, the author decided on to just focus on featured articles as they would be already vetted as high quality. The dataset that remained contained 1554 featured articles and 9513 random articles for the control, and was split 67% for training vs testing.

For the results, the binary classification test concluded with a 96% accuracy when computing between if the article had more or less than 2000 words. More complex classification functions such as k-nearest neighbors and random-forest classifiers all achieved an average of mid 90s. This was 10% better than more complex metrics that previous authors used.

In terms of our thoughts on this paper, we feel that despite the compelling case that makes word count a great measure of quality, we feel that it lacks some edge cases to make it work all the time. An article does not have to have a lot of words to be great, especially in Wikipedia's case as it is meant to report factual information, not share opinions. Hence, an article can contain less than 2000 words but still have the quality of those articles that are extremely long. Also, by making word count the only metric of quality, there is a chance that editors on the platform may abuse this metric and expand articles for no reason. Readers like brevity, and by prolonging articles just for the sake of improving the quality will hurt the end reader as they are more likely to skim through the article rather than carefully read each word. Therefore, despite the positive results that came out of this paper, there are some consequences with this metric.

2.4 Wikipedia network analysis of cancer interactions and world influence

The next paper that we will look at was written in 2019 by Guillaume Rollin, José Lages, and Dima L. Shepelyansky. In it, the authors explore the overlap between Wikipedia articles on types of cancers and the fatality rate of them. They utilize the Google PageRank algorithm to sort the articles of cancer and try to compare it to the WHO's diseases study. They also explore cancer drugs and world countries to see overlaps between the types of cancer and the types of drugs that are used to treat them.

In the introduction, the authors speak about how cancer affects 1 out of 6 people in the world, and go over the details on how it is expected to increase. The section continues onwards to speak about the objectives and experiments that the authors are going to run to create this network. Rollin et al. [2019] explores how Wikipedia articles can create links between cancer types, and how to extract information from the network that may help researchers find similarities between them. The authors used the

Google matrix and PageRank algorithms to find the most popular cancers on the site and cross-reference them with the deadliest cancers in the world. The introduction is also where the researchers outline that the network is made up of 37 cancer articles and 203 cancer drugs. They also selected 195 countries to see the impact these cancers have in these regions.

The description of data sets and methods section explores the network that Rollin et al. [2019] built and describes the various algorithms used to find the connections between Wikipedia and the real world. The Wikipedia network had 435 nodes made up of country names, the cancer types, and the cancer drugs. The Google matrix algorithm was constructed with an adjacency matrix made up of all the nodes that we discussed prior. The influence of each node was calculated using the normal PageRank algorithm. The reduced Google matrix algorithm allowed the researchers to determine the sensitivity of the PageRank algorithm as it allowed them to find specific links between each node.

In the results section, Rollin et al. [2019] investigated the cancer distribution using the PageRank algorithm. This resulted in lung and breast cancers being the highest occurrences. For the cancer drugs, Talc, Methotrexate, and Thalidomide were all at the top three. This made sense as they are used to treat a wide range of cancers. The outcome of the PageRank suggested that the Wikipedia network had the same influence globally when compared to cancer and diseases information. The next set of results is the comparison between the Wikipedia network and the GLOBOCAN (GBD) study on cancer significance. More than 70% of top nodes in the Wikipedia network matched the GBD study and the network was reliable enough to predict the most fatal cancers in the world. The network also reached 80% resemblance for the top 5 cancer types. The reduced Google network was able to find hidden links between the most devastating cancers and it formed clusters of similar cancer types. Lung cancer was also mentioned by all cancers except 2. They then introduced countries into this reduced network, and it led to more results. Cancers that are found in the digestive system tend to be from countries in Asia, such as liver cancer that impacts Eastern Asia the most. Colorectal cancer had stronger connections to western countries while blood cancers were more found in African countries. The clusters of cancer types that formed with these hidden links resulted in cancers that tend to affect the same body functions grouping together, further enhancing the interesting nature of the reduce network algorithm. When the researchers combined the cancer drugs to the reduced network, they found that the drugs that tend to treat those types of cancers matched with the cancer that they are made to treat. Also, connected cancers in the same cluster also relied on the same drugs.

Rollin et al. [2019] concludes the paper by illustrating how accurate the reduced Google matrix is when trying to create a network that tries to connect multiple related elements. The hidden links that were found between each cancer, country, and drug allowed the team to further investigate how Wikipedia's network of articles can predict clusters and connections between various unrelated objects. It is interesting to see how something like a community powered online free encyclopedia can create networks that can accurately model the real world. The other takeaway from this article is that Google has definitely altered the way that information bubbles up with their PageRank algorithm. Their reduced matrix now can also allow us to build better networks and provide more personalized content.

2.5 Measuring article quality in Wikipedia: Models and evaluation

The fifth related paper that we will be covering is written by Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lau, and Ba-Quy Vuong, all a part of the Nanyang Technological University in Singapore. This paper was published in 2007, and a part of the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. The authors explore why Wikipedia is never trusted by experts and present three new measurement models that could introduce the metric of quality into articles. Those included the Basic, PeerReview, and ProbReview models.

In the introduction and related works sections, Hu et al. [2007] goes into detail about their motivations. The authors talk about the issues with how unreliable articles can be on Wikipedia and therefore, making it hard for the general public to use it as a trusted source. They wanted to allow users to know which articles are high quality work and alert Wikipedia contributors of articles that are low quality. This could also improve navigation in the site and make users find articles more easily. Of course there are issues such as that there are many articles within Wikipedia, there are numerous subject matters, different editor skill levels, and the threat of spam and abuses on the site. The researchers outline their three models based on both the articles and the editors, and then create experiments to test their models on. In the related works, Hu et al. [2007] hit on previous papers that explored this topic and spoke about how previous authors went through how citations, number of edits, and number of unique editors all play a part in improving the quality of articles.

In section 3, Hu et al. [2007] breaks down their three models that they have created to determine quality within Wikipedia articles. Those include Basic, PeerReview, and ProbReview. The Basic model is dependent on the authority score of each editor. This means that the higher the authority for that editor, the better quality that article will be. This of course is dependent on the entire edit history for that writer. The PeerReview model is based on if edits remain over time, that edit is good quality. That is also true if that edit was made with an editor with a higher authority score. Finally, the ProbReview model looks at an edge case that the PeerReview model glances over. That is, the editor that writes and edits the article might not review the entire article again before publication. This means that they might miss something that has been added to the article at a later date and therefore, the quality may decrease. Therefore, ProbeReview tries to determine what the editor that had the high authority score has read and proofread. This is dependent on when and where last that user has altered text and how many more revisions have occurred since then. The researchers also define Naive as the baseline model to compare their results with. They finally list pseudocode on how they would go upon computing these models.

We then reach the experiments section of the paper. The dataset consisted of 242 articles which already had manual class labels already assigned, indicating the quality of work. Some of them were featured articles or the best, a-class, good, etc. Then, the data was cleaned and the Wikipedia API was used to collect all of the revisions made in each article. This allowed the researchers to get over 100,000 unique users and only 29% of them were registered. The two metrics used to determine the accuracy of the data was the Normalized Discounted Cumulative Gain at top k (NDCG) and Spearman's rank correlation. NDCG is useful to evaluate articles that are ranked, especially in this case when there are numerous quality rankings given to each. 1 represents great performance while 0 represents poor performance. Spearman's rank is useful for comparing the same object with different rankings. It uses a similar output number as NDCG, however the score can reach -1.

Onto the results, for NDCG, the PeerReview and ProbeReview models outperformed the baseline Naive model. However, the Basic model did almost the same as the benchmark model. Hu et al. [2007] also points out that when K increases in value, the models do improve their NDCG scores. The outcome from this test was that the performance was better with these models than trusting the length of the article to be representative of the quality of it. For the Spearman's rank correlation, the ProbReview model performed the best and that kept to the theme that was witnessed from the NDCG test. Hu et al. [2007] then changed around the parameters that were being used by the model and also incorporated the article length into the calculations to see if this value improves or decreases the quality of the articles. When doing this, the baseline model performed better than the Basic model. This means that article length does in fact increase the quality of the article. Lastly in the results sections, the authors talk about the authority scores of users and how registered users would have higher scores than unregistered users. They then concluded that user interactions do not impact the quality of the work and should not be used as an indicator moving forward. However, they were surprised that article length improved their scores.

Overall, this paper was interesting as it dealt with a more sophisticated approach to calculate the quality of text. We really enjoyed the parts on authority scores of the users and we feel that this is a great way to class users based on their contributions. As readers of Wikipedia, writers that have their authority scores listed would make it easier to trust the article that is being read.

2.6 Network analysis of user generated content quality in Wikipedia

In this sixth research paper by Myshkin Ingawale, Amitava Dutta, Rahul Roy, and Priya Seetharaman, the authors investigate the idea that there may be a relationship between the quality of user generated content, and the contributors which generate and interact with that content. To that end, they use data from Wikipedia which has a discussion and consensus based content creation process where contributors decide through conversation and consensus building in order to collaboratively determine what appears in articles. Such is the basis of the high level of accuracy and reliability of Wikipedia articles compared to traditional encyclopedias. Specifically, the authors want to investigate whether structural holes and centrality are related to the quality of output.

To do that, they define high quality as what Wikipedia itself constitutes as a "perfect article", which means that it includes being from a neutral point-of-view, is verifiable, and has no original research. Additionally, they also have their own method of determining high and low quality articles which they conclude that their ranking on search engines are "at least appropriate" [11]. Wikipedia also has a system of Featured Articles, which are articles voted on to be perfect articles according to the requirements by Wikipedia, which gives a clear example of what all articles should aspire to be like.

Now that they have defined what quality is in a Wikipedia article, they then look for structural holes with the clustering coefficient and average pathlength metrics from the network structure. The clustering coefficient measures how well connected a node is to its neighbors, and the average pathlength measures how central a node is, determined by how it is connected to every other node in the network. Their dataset was made up of Wikipedia's revision history files from six different languages. The result of their analysis is that a node's access to resources is affected by its location, and that increased centrality also increases the quality of output. Additionally, featured articles will appear to span structural holes, which means that those high quality articles connect networks that would otherwise have been sparsely connected. As for the user network, relatively speaking, only a small number of users actually make up most of the contributions.

As a result of their findings, they have determined that article hubs which span structural holes connect different networks together, and due to that have an unusually high amount of contributions that are not redundant. That also makes it likely that those hubs will likely be featured articles. As for users, that limited number of users which make up most of the contributions should be nurtured by organizations. Their work also supports the knowledge network theory with the structure hole metaphor in showing that the network brokerage mechanism is seen clearly through the quality of Wikipedia articles.

Some of the limitations they encountered were that during the conversion from the affiliation network to the user and article networks, some data was lost. Another limitation they faced was that they were not able to determine the impact of each individual revision, as some of them would naturally have more of an effect on the quality of the article. Furthermore, they acknowledged that they only used six of the best language Wikipedias, and so the results of their research may not extend to all languages used for Wikipedia. Future work they see as a result of their study are to look further into how collaborative creation works and their social systems. There are also many other social networks that could be analyzed such as Facebook and LinkedIn that could help in the research of social interactions with technology.

This paper has shown an interesting aspect of Wikipedia article quality, and their correlation to user contribution. The analysis of user generated content was relatively new, and it reinforced the idea that nodes which connect networks obtain numerous benefits from their position in the network. We see this paper as insight into some of the early work into the analysis of networks in Wikipedia, and hope to perform a somewhat similar analysis of Wikipedia networks, but from a different perspective.

2.7 World Influence of Infectious Diseases from Wikipedia Network Analysis

The seventh paper we take a look at is by Guillaume Rollin, José Lages, and Dima L. Shepelyansky. Through this paper, they look into the influence of infectious diseases on countries based on their sensitivity using both the Google matrix and the more recent reduced Google matrix analysis methods.

The use of the Google matrix allows them to rank their data into PageRank and CheiRank. PageRank ranks the Wikipedia pages according to the probability of a random surfer jumping to it, while CheiRank is the inversion of PageRank, and so it ranks the pages on how communicative it is, or how many outgoing links it has. Their choice of using the reduced Google matrix analysis method, or REGOMAX has to do with its unique ability to not only be able to connect direct links, but also to be able to infer indirect links. These indirect links provide valuable insight into interactions that would have otherwise been overlooked. With REGOMAX, they also analyze the PageRank sensitivity relative to a disease.

For the results of their analysis, they begin by discussing the network of direct links. Using REGOMAX, they had 195 countries, and 230 infectious diseases analyzed, resulting in 425 nodes. These were visually represented using the Cytoscape software, as seen in their paper.

This network of direct links is also represented in the adjacency matrix which represents links as white dots, and are grouped by countries and types of diseases. Through that, it shows different densities of links, which are especially visible amongst the same subgroups.

Next, they look into the PageRank and CheiRank results. For countries, they find that as expected, the countries with the top three PageRanks in order are the United States, France, and Germany. As for diseases, the top three PageRank in order are Tuberculosis, HIV/AIDS, and Malaria. Those results show infectious diseases which have spread worldwide at top positions. The top CheiRank disease was also indicated to be Burkholderia because of its many outward links.

For further analysis, they look at the three components which make up the reduced Google matrix. Those components are G_{pr}, G_{qr}, and G_{rr}. G_{pr} doesn't provide that much information, as it is similar to the PageRank columns, but G_{qr} and G_{rr} have similar weights, which indicate that the direct and indirect links have similar amounts of contribution.

In order to create a friendship network, they use the process of taking the group leader of each disease group to use as the nodes to start with, and then at each iteration, the two best friends of a leader are used to create two new nodes. After that, two new best friends from each of the previous nodes are added, and so on until no new friends can be added. This then creates the network.

That network shows some important interactions between diseases such as the Sepsis's interactions with Pneumonia and Malaria at a first level link, and between Meningitis and Sepsis at a second level link. The interactions Sepsis has with Pneumonia and Malaria has to do with it being a potential symptom of Malaria, as well as a complication with Pneumonia. As for the interactions between Meningitis and Sepsis, they create a closed loop as patients with Meningitis are likely to also develop Sepsis early on. The red arrows show indirect links such as between Malaria and Desmodesmus, which do not directly link to each other, but occur because they are both still waterborne diseases.

Additionally, the sensitivity of countries' PageRank to different infectious diseases were also measured to find out how different countries are influenced by infectious diseases. With the REGOMAX analysis, they were also able to uncover indirect interactions, such as that the PageRank sensitivity of countries like Rwanda and the Marshall Islands are influenced by HIV/AIDS links with the United States despite not being directly connected to it.

Comparing their results to the World Health Organization's, they find that their rankings mostly overlap for the highest ranking diseases, and still around 50% for all the rankings combined. However, this generally does confirm though that Wikipedia network analysis can provide reliable information.

This paper has shown us that the analysis of Wikipedia network can have its merits, as it has come a long way to supersede the traditional Encyclopedia. Being an online resource, it also allows us better accessibility, and thus can not only be used for infectious diseases as this paper covered, but also for many other topics.

2.8 Using big data & network analysis to understand Wikipedia article quality

For the eighth paper we look at by Jun Liu and Sudha Ram, we take a look at the quality of Wikipedia articles and how they are influenced by social capital in the forms of external bridging, functional diversity, and internal bonding. Their goal in this research is to understand the relationship between the contributors on Wikipedia and the quality of articles, as the internet breaks down that old idea that collaborative development can be detrimental if there are too many authors. They found that prior research shows that Wikipedia article quality isn't simply a matter of the number of edits or editors, but related to social capital, and so look into that impact on the quality of articles.

In their literature review, they look at other research on what affected the quality of Wikipedia articles. Some factors they found are that the balance of male and female contributors affected the quality, as well as the kind of editors which make edits. With a poor male to female balance, the comprehensiveness of an article may suffer, and all-round editors often result in higher quality articles. As for a social network's effect on article quality, other research showed that interactions between co-editors likely results in higher quality articles. The role of the researchers in this paper is to understand how the forming of relationships as a result of working on the same article, or an editor's position amongst the network of editors affects article quality.

Onto their hypotheses and theory for social capital, they begin with explaining that social capital is the idea that people who are better connected will likely perform better even if the skills and abilities of people were the same. Social capital is further split into three categories. One is bonding social capital, which are ties amongst members of the same group. Another is bridging social capital, which are ties between groups to other groups. The third is functional diversity, which is how different contributors affect articles in different ways according to their roles. Each of these have their own effects on article quality.

Starting with the effects of internal bonding, it can be thought of as similar to network closure. In social networks, internal bonding can affect a team by improving the flow of information, having more of a consensus, and having a shared understanding of problems within the group. As for Wikipedia teams, internal bonding has the effects of trust between members, improved coordination, and the maintenance of resources which affect the quality of articles. Homophily is also expected to have a role in internal bonding, and so teams are likely to have similar interests, and thus the social capital accumulates as the links among nodes in the network continues to grow. So through internal bonding, they have three hypotheses. These are that internal collaborations, external collaborations, and homophily of contributors all positively impact the quality of Wikipedia articles.

As for the effects of external bridging, they are the ties to members outside of a team. This allows new knowledge and ideas to come from people outside of a team, and improves the overall performance of the team. Furthermore, people who connect teams together fill structural holes and thus play the role of a broker for a team to other teams. In terms of Wikipedia article quality, they found that external bridging can affect it in three different ways. They are that it allows article information to be verifiable, ensures that contributors have multiple points of view, and allows a diverse source of knowledge in contributing to the quality of articles. Thus, their hypothesis for external bridging is that more structural holes spanned by contributors will positively impact the quality of Wikipedia articles.

Additionally, they find that internal bonding and external bridging can complement each other, and have a multiplicative effect in the quality of articles. So in this manner, they have three more hypotheses. They are that interactions between internal collaborations, external collaborations, and homophily with spanned structural holes all positively impact the quality of Wikipedia articles.

Lastly, they also look at functional diversity, which is the effects on article quality depending on the diversity of roles within a team. In general, the idea is that if there are enough people ensuring the quality of an article, then its quality will naturally increase. In the case of Wikipedia, the way articles are written, and the diversity of edits can both affect quality. So their hypothesis is that functional diversity in a team will positively impact the quality of Wikipedia articles.

Next, they collected data to test their hypotheses. The data they collected include articles' quality ratings which from lowest to highest are Stub, Start, C-class, B-class, Good Articles (GA), A-class, and Featured Articles (FA) [9]. The FA and GA statuses are determined by consensus by the contributors and some delegates which give the final say. As for the class system, they are managed by WikiProjects. The other set of data they collected is how contributors edited articles, and the history of edits. Then for the construction of the social network, they set nodes as the contributors, and the edges as the same sentences people have worked on together, with the weight being the number of the same sentences.

As a result of their analysis, they found that their results were consistent with the hypotheses, and that social capital does play a role in the quality of Wikipedia articles. Their contributions include developing the theory to determine Wikipedia article quality, the techniques to compare how two different networks can affect the same environment, how social capital can be extracted from Wikipedia, the importance of single mode networks, and defining the importance of functional diversity in terms of role distribution.

Their study has shown us that Wikipedia article quality is affected by social capital, among many other factors we looked at above. It also confirms how collaboration can generate quality content even in a seemingly anarchic environment and will continue to, as the world becomes more closely connected with the internet. Apart from Wikipedia, this study also has implications in the real world, as social capital and collaboration also exists outside of Wikipedia articles.

2.9 Interactions of pharmaceutical companies with world countries, cancers and rare diseases from Wikipedia network analysis

Now for the ninth paper, the authors are Guillaume Rollin, José Lages, Tatiana S. Serebriyskaya, and Dima L. Shepelyansky. Through this paper they analyze Wikipedia articles with links between pharmaceutical companies, countries, rare renal diseases, and types of cancers. As Wikipedia has become larger and more widespread, it has become an increasingly accurate source of information especially in regards to scientific topics that are actively maintained. As such it has become a useful dataset, and the authors aim to use it to analyze the influence and mutual interactions in Wikipedia between the world's 34 largest biotechnology and pharmaceutical companies, as well as their relation to 195 countries, and their relation to 47 rare renal diseases, and 37 types of cancer. Their hope is that through this study, they will be able to better understand how different pharmaceutical companies may specialize in curing different diseases.

Their methods of analysis include using the Google matrix and reduced Google matrix (REGOMAX) methods. With the Google matrix, they are able to rank nodes by PageRank, and CheiRank. PageRank ranks the nodes by centrality and influence, while CheiRank ranks the nodes by diffusivity and outgoing edges. As for the REGOMAX method, it is similar to the Google matrix, but with more of a focus on a subset of nodes, and is a combination of G_{rr}, G_{pr}, and G_{qr}. G_{rr} is the submatrix of nodes that are deemed as important, but still have the node information as from the Google matrix. G_{pr} is where all columns are the same PageRank vectors as sorted by the Google matrix analysis. G_{rr} are the indirect links between nodes where connections occur between nodes within the submatrix, but happen through nodes outside of the submatrix. With REGOMAX, they are also able to identify the sensitivity of the PageRank, given changes to other nodes that affect it.

The datasets they use include all the articles in the English Wikipedia as of May 2017 which is 5416537 articles, and the hyperlinks between articles, which is 122232932 hyperlinks. Those are used for the Google matrix analysis. For the reduced Google matrix analysis, they used a selected group of 195 countries, 47 rare renal diseases, and 47 types of cancer. For REGOMAX, that makes up 313 Wikipedia articles. With this data, they analyzed the PageRank to CheiRank distribution to show that articles on pharmaceutical companies have influence and diffusivity that are correlated. They also ranked the market capitalization and largest capitalization for each company, which showed that the economic performance of these companies also had an impact on their influence. Some irregularities they did note however were that older Wikipedia articles tended to have higher PageRank and CheiRank indexes as they had more time to accumulate links, but that does not extend to articles for countries since they were mostly created around the same time.

Onto their analysis of the REGOMAX, they have many findings in regards to direct and indirect links among the pharmaceutical companies and rare renal diseases in the visual representation. With pharmaceutical companies, they find that there are many direct links between each other, showing the competitive nature of the industry in such a way that companies are often related to each other. They also find that one disease, the Fabry disease, had not been directly cited by one of the larger pharmaceutical companies, and thus has been given the status of an orphan, as its only direct link is through Shire, which manufactures Replagal. One of the indirect links which they found were between Abbott and Johnson & Johnson by going through Advanced Medical Optics. The hidden link is due to Advanced Medical Optics being acquired by Johnson & Johnson, but having previously been owned by Abbott.

Moreover, they made network structures from the reduced Google matrix, as can be seen in the figure in their paper. In that figure, the top five pharmaceutical companies are first represented with purple circles with black dots in the center. Every iteration, the two best connected pharmaceutical companies and countries are added, with new pharmaceutical companies having just purple circles, and this goes on for three iterations. The black arrows indicate direct links, and red arrows indicate indirect links. Depending on the iteration, the form of the arrow will change as well. From this they find that Pfizer's company friends are Johnson & Johnson and GlaxoSmithKline. Pfizer's country friends are Ireland and Italy. These links are easily identifiable and understandable by following the hyperlinks which link them together. As for indirect links, there are many of them pointing towards the United States, and is an interesting indication that most pharmaceutical companies are from the United States.

Sensitivity of countries to pharmaceutical companies and rare renal diseases were also analyzed. Some interesting insights they extracted include that Italy is sensitive to Pfizer due to the historical fact that Italy provided citric acid to Pfizer until World War I. Also, countries like Germany are sensitive to the Kallmann syndrome due to the fact that the person who described the disease was known as Franz Josef Kallmann from Germany.

In general, this paper successfully used Wikipedia networks to analyze interactions between groups of related topics. We find that it is interesting how many metrics they were able to gather just from the interactions between them, and it shows the value of Wikipedia network analysis.

2.10 Do editors or articles drive collaboration?

Finally for the last paper we will cover, the authors are Brian Keegan, Darren Gergle, and Noshir Contractor. Through this paper, they consider how editors and articles interact to determine if these interactions affect collaboration. More specifically, they analyzed the difference in collaboration with breaking news compared to articles that are just similar. They also analyzed the different features of editors and articles that may affect the way collaboration occurs. Through their analysis, they also share how to use p*/exponential random graph models (p*/ERGMs) to perform a multi-level network analysis, and how different levels of editors and articles can affect the collaboration structure.

Background information on the topic shows that Wikipedia is interesting in that it ensures that editors are motivated to continue editing despite the fact that most users don’t actually do any editing. Thus they claim that editors traits, articles features, interactions between editors with article features, and interactions between articles with editor traits will explain how Wikipedia is self-organized.

For studies focused on articles’ impact on collaboration, they share that too many authors can be detrimental to the quality of Wikipedia articles. However, the authors argue that without understanding the interaction between editors and authors, there may not come an understanding of how breaking news can simultaneously have high quality and a high number of editors. Editors have varying goals and levels of qualifications depending on the article attribute as well. So the authors have two hypotheses. The first is that article attributes affect the number of editors, such as the difference between breaking news and non-breaking news articles. The second is that article attributes also affect the types of editors who are attracted, and similar types of editors are more interested in working with each other.

For studies focused on editors’ impact on collaboration, an editor’s role and experience has seen both potentially impact their levels of collaboration. Such roles could include editors who only edit breaking news articles, or editors who can be considered caretakers because they have a lot of experience, and have edited many articles in a particular field. Thus, the authors have suggested two hypotheses. The first is that an editor’s attributes will affect how many articles they will edit. The second is that an editor’s attributes will also affect what kinds of articles they decide to edit.

Other explanations for the variations in the collaboration structure were also discussed. These include factors such as how related an article is to an editor, such as geographic distance to the topic, and the time an editor has spent in the community. The authors have considered these factors by including them in their variables of interest.

Their approach for analysis is to use the multi-level statistical model of p*/ERGM. It will allow them to understand dependencies on both endogenous and exogenous tendencies such that they can test their hypotheses. Endogenous tendencies would refer to local structures that an article’s level of collaboration. While exogenous tendencies would refer to factors outside of the local network, such as experienced editors attracting other experienced editors to collaborate.

Their data includes the data on articles, editors, and edits. With that they were able to create 23903 edges between articles and editors. For modeling this, they use a bipartite method, to combine the interactions of both editors to articles, and articles to editors. The result of their analysis shows that breaking news articles will attract more editors than other articles, which proves their first hypothesis that article attributes can affect the amount of editors. However they also find that unlike their third hypothesis that editor attributes will affect the amount of articles edited, as they see that more experienced editors actually edit less than less experienced ones. For their second hypothesis of article attributes affecting the types of editors who work together, their findings show that it is true, as editors with similar experience tend to work together. As for their fourth hypothesis that editor attributes will affect the types of articles they edit, as they found that an editor’s experience will influence them to edit similar types of articles.

As a result of their study, they’ve understood the question of how breaking news articles, despite experiencing a chaotic environment of editors, can remain high quality. This has to do with editor attributes, and since they will likely work on similar articles, it will assist in the quality of any of the articles they work on due to their experience. Their study’s implication is also that they showed how to use, and the value of the p*/ERGM methods so that many new analyses can be made. Though, they do admit that p*/ERGM does have a high computational requirement. Still, they hope that there will be more use with the method to do more analyses.

We find that this study has shared valuable insights into how collaboration occurs in a Wikipedia article. Understanding this has shown us that the level of collaboration can vary depending on a number of factors, and that the editor to article relationship is not to be underestimated in its effects on a Wikipedia article.

3. Implementation

The network that we built to study Wikipedia further is based on Wikispeedia data [12]. This is a web game that challenges players to navigate to a certain destination (article) on Wikipedia from a given article. From the user data, Robert West and Jure Leskovec devised a dataset containing all of the links that players clicked on [13]. We downloaded the database from the SNAP website and extracted the files [12]. Once done, we opened the file in Python and iterated through the links file that included the directed edge between source and destination nodes. After making sure that we were not including comments, and formatting the text using the urllib library, we used NetworkX to construct the network. We have imported all of the data into a NetworkX graph object which will help us run methods that can help us understand the link network of Wikipedia. NetworkX is a powerful Python library that includes multiple built in functions that allows us to gather key statistics from the network and enables us to import data into matplotlib, a Python graphing library, to visualize the data in the graph. We will use the categories.tsv file that holds the category attribute for each article. This will help us in the community detection part of this paper, as it will provide us more information about the nodes in each community. In the results section below, we will look at more simple statistics such as the degree distribution and which nodes have the highest degrees. We will also try to visualize the data using graphs. Once we report these numbers, we can then move to our goals of trying to find connections and communities between nodes.

4. Results

In this section, we will explore the statistics of the network from Wikispeedia and then outline articles that have the highest values in three centrality categories. We then will report our findings of the community assessment.

4.1 Network Statistics

Starting off with the more mundane statistics, there are 8722 nodes or articles in the graph and 119882 edges (links). All edges are weighted the same, as all links on Wikipedia are the same priority. The edge density of the graph is 0.00157. The network has an average degree distribution of 27.49. In figure 1, we have graphed the degree distribution on a log scale and in figure BLANK, we have graphed it on a regular scale without normalizing the values.

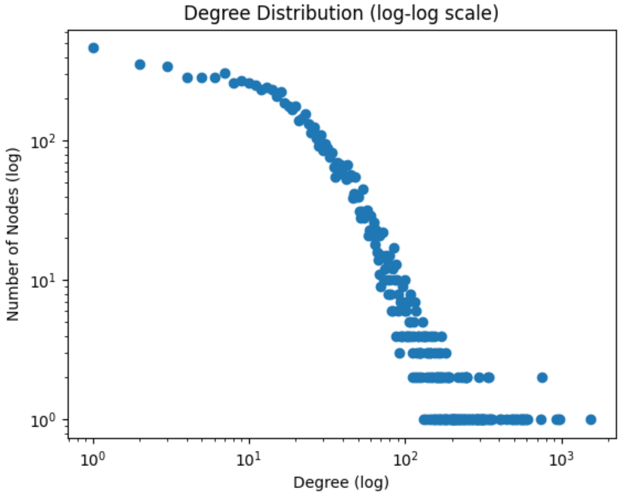


Figure 1**: The degree distribution of all nodes, based on a log scale with normalized values.**

As we can see, the graph follows a power law or scale free distribution. This signifies that there are numerous large hubs in the network and these hubs have many in and out edges.

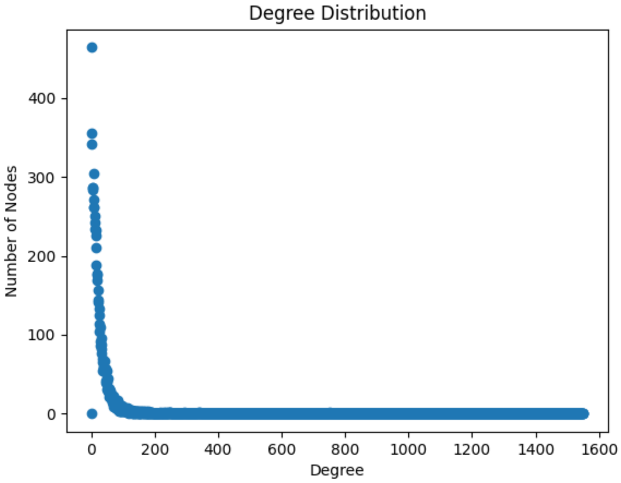


Figure 2**: The degree distribution of all nodes, without normalization and graphed on a normal scale.**

Again, the log free graph demonstrates a power law distribution with some few nodes having a large degree.

4.2. Notions of Centrality

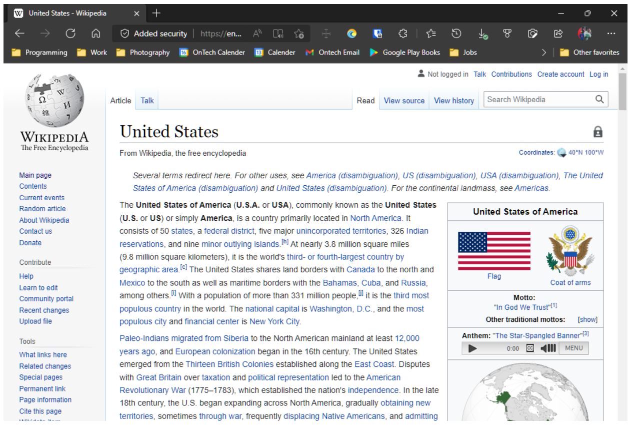
We will list the top 5 nodes (articles) in the network with regards to their degree, out-degree, and PageRank values, and then theorize why these nodes have the highest values in their respective category.

4.2.1 Degree centrality: The degree centrality is the first measure that we will explore. In table 1, we have outlined the article name and the degree value of that article.

Name	*United States*	*United Kingdom*	*France*	*Europe*	*World War II*
Value	0.1778	0.1115	0.1099	0.1069	0.0861

Table 1: **This table contains the top 5 articles with the highest normalized degree values in the network.**

As we can see, the top 4 of 5 articles with the highest degrees are countries or continents. These articles tend to have a lot of links linking topics together and also have a lot of news stories referring to where certain stories took place. We can see the United States Wikipedia article in figure 3. Each of the blue text links are links to other Wikipedia articles and that encapsulates how vast these articles are. The last article that is listed in the top 5 of table 1 is World War II. World War II is still one of the most historical moments of human history. The article covers a wide range of topics, from each of the significant battles that were fought, the events leading up to, proceeding, impacts, and more. This included a wide range of topics, linking to a wide variety of affects, while also being a hub of inbound traffic, as there are many policies that stem from the war.



United States

From Wikipedia, the free encyclopedia

Several terms redirect here. For other uses, see America (disambiguation), US (disambiguation), USA (disambiguation), The United States of America (disambiguation) and United States (disambiguation). For the continental landmass, see Americas.

The **United States of America** (**U.S.A.** or **USA**), commonly known as the **United States** (**U.S.** or **US**) or simply **America**, is a country primarily located in North America. It consists of 50 states, a federal district, five major unincorporated territories, 326 Indian reservations, and nine minor outlying islands.[h] At nearly 3.8 million square miles (9.8 million square kilometers), it is the world's third- or fourth-largest country by geographic area.[c] The United States shares land borders with Canada to the north and Mexico to the south as well as maritime borders with the Bahamas, Cuba, and Russia, among others.[i] With a population of more than 331 million people,[j] it is the third most populous country in the world. The national capital is Washington, D.C., and the most populous city and financial center is New York City.

Paleo-Indians migrated from Siberia to the North American mainland at least 12,000 years ago, and European colonization began in the 16th century. The United States emerged from the Thirteen British Colonies established along the East Coast. Disputes with Great Britain over taxation and political representation led to the American Revolutionary War (1775–1783), which established the nation's independence. In the late 18th century, the U.S. began expanding across North America, gradually obtaining new territories, sometimes through war, frequently displacing Native Americans, and admitting

Motto: "In God We Trust"[1]
Other traditional mottos: [show]

Figure 3**: The Wikipedia article for the United States.**

4.2.2 Out-degree centrality: The out-degree centrality is the second measure that we will use to find the hubs in the network. In table 2, we have outlined the article name and the normalized out-degree value of the articles.

Article Name	*Out-degree Value*
United States	0.03371
Driving on the left or right	0.02923
List of Countries	0.02797
List of Circulating Currencies	0.02706
List of Sovereign States	0.02476

Table 2: **This table contains the top 5 articles with the highest normalized out-degree values in the network.**

Looking at the out-degree top 5, a pattern has formed. That is that 3 out of the 5 nodes are lists. As lists on Wikipedia tend to include a lot of internal links, they would be on top of the pile when it comes to the amount of links out. "Driving on the left or right," includes a lot of links to countries that have their own articles. The United States article again tops the pack and that is not surprising as the United States has a lot of links to a wide variety of topics. From technology, laws, wars, and more, the chance of this article having the highest number of outbound links is high and it is proven by this network. Wikipedia not only has stories of events, but also contains lists and other rankings that help finding information easier. Wikipedia is now so much more than just a regular encyclopedia, and that is proven at how diverse the articles are.

4.2.3 PageRank centrality: The PageRank centrality is another interesting measure that we will explore. In table 3, we have outlined the top 5 nodes and the PageRank value of each.

Name	*United States*	*United Kingdom*	*Scientific Classification*	*Europe*	*England*
Value	0.005275	0.002871	0.002699	0.002638	0.002556

Table 3: **This table contains the top 5 articles with the highest PageRank values in the network.**

The last centrality measure that we calculated is PageRank, a classic way to determine the worth of pages. As we want to see the most influential articles on Wikipedia, this is a great way to determine this. As expected, the United States article is number one, and 4 of the top 5 PageRank articles are countries and continents. As explained earlier, these articles tend to have more links, therefore the article branches out more further than regular articles. Furthermore, this also tells us that these 5 pages are the most important pages in Wikipedia.

Other centeralities were considered when choosing these three measures, such as closeness and betweenness. However, the values and articles were very similar to values already in the previous 3 tables. Therefore, we felt that we have already discussed above why various articles would in turn be in the top 5 of those tables as well.

4.3 Community Assessment

We will now explore the communities that are found in the network. As there are over 8000 nodes, we will be using the category attribute for each node so that it would make it easier for us to spot irregularities in each community. However, there were many nodes in the dataset that did not have a selected category. Hence, a portion of each community was not defined, making it harder for us to fully gather the entire population of nodes. As communities one to three have over 2500 nodes, it was impossible to go through all undefined nodes in our timeframe and add a category to them, after determining what

	# of Nodes	*No Category*	*Science*	*Geo-graphy*	*Math*	*People*	*Reli-gion*	*Lang-uage*	*History*	*Busi-ness*	*Tech-nology*	*Every-day Life*	*Citizen-ship*
Group 1	3182	1495	**979**	208	43	115	43	27	49	18	53	78	54
Group 2	2638	1193	53	**621**	1	198	48	55	221	26	47	71	81
Group 3	2621	1292	34	200	1	**340**	29	98	125	19	187	157	52
Group 4	79	**25**	4	1	0	15	4	1	19	0	10	0	0
Group 5	48	9	17	1	0	0	0	0	0	0	0	**21**	0
Group 6	39	4	6	4	0	0	0	0	7	**17**	0	1	0
Group 7	26	3	0	7	0	2	**9**	1	0	0	4	0	0
Group 8	23	3	0	0	0	0	1	0	0	5	0	0	**14**
Group 9	11	2	1	0	0	1	0	1	0	**2**	0	1	**2**
Group 10	10	0	0	**10**	0	0	0	0	0	0	0	0	0

Table 4: **The community table showing the number of articles for each group (community).**

the article is about. We ran the greedy_modularity_communities NetworkX method and then iterated over the 23 communities. The results were written to two text files, one for the article names and one for the category. We used the search tool in Excel to find the number of articles in each subject matter in each community. Table 4 breaks down the 10 communities that have been calculated. The bold numbers in each row represent the highest category amount within that community. We felt that any communities lower than 10 nodes were irrelevant as there is not enough data to understand what that community is made up of. For the table, we used the categories that were already created in the dataset and combined a couple that we felt were similar. We combined the "countries" category with the "geography" category. Also, we combined the "design and technology" category with the "I.T" category. The "citizenship" category dealt with politics, uprisings, and law. The "everyday life category" was used for sports, lifestyle, and entertainment events. It also included food. The "people" category included the articles about a person, regardless of which field they are in. We felt that categories with below 20 total articles were irrelevant.

Moving to the analysis of the communities, we can see that the majority of the nodes are in the first three categories. They account for 96.6% of all nodes. In them, we can clearly see that group one included a lot of science and geography articles. In fact, 91% of all science defined articles were located in the first group. The math articles also were abundant in group 1, in fact it contained 43/45 articles across the dataset in community 1. We can definitely classify group 1 as the science and math community, and as these two categories rely on one another, it makes sense that they would be bunched up in the same group. Group 2 included a lot of geography articles. However, geography articles were placed quite evenly in groups 1 and 3, indicating that geography articles tend to branch out much further than other genres. The history, citizenship, business, religion, and geography categorized articles peaked in group 2, when compared to any other community. These genres tend to be related to each other and it further emphasizes the correctness of the community algorithm. Group 3 contained the most everyday life and technology articles, illustrating to us that technology has impacted our lives so much that it ties into what we do everyday. When looking at the smaller communities, there are much more specified areas of articles when compared to the top three groups. For example, group 5 included a lot of food and herbs articles, making the everyday life and science categories account for 38 out of 39 articles in that group. Group 6 included a lot of articles on stocks, financing, and other business related topics. Group 8 focussed almost solely on rights movements, such as women's rights protests and other notable rallys. All of group 10's articles were to do with past and active volcanoes. The community assessment allowed us to understand how various categories of articles interact with each other on the free, open encyclopedia.

5. Discussion

Looking back at the 10 research papers that we summarized in the related papers section, we definitely could plot similarities between each of the articles. For starters, most of them dealt with trying to add quality as a metric to Wikipedia articles. This allowed readers to trust the platform more than they have done so in the past. The ways the researchers went upon doing this however were all different. Some merely just focused on article length [8], while others looked at edit longevity, while a bunch of other metrics [2] [9] [10] [11]. All of these articles created the metric and experiments, and then tested it using a wide range of use cases. All of them are successful, however which one will Wikipedia actually use? None so far at this very moment. That goes to show that scientific articles can continue to theorize ways to improve Wikipedia with quality scores, but will Wikipedia actually implement these scores moving forward? As students, we are told from a young age that Wikipedia is not a trustworthy source. It has edits that are not true and it has not been peer reviewed. Does that make Wikipedia a fictitious site? In our eyes, no but for Wikipedia to be a trustworthy source, it has to implement ways to improve the quality of articles, just like some of the papers that we looked at suggest. Despite this lack of trustworthiness, Wikipedia has grown into a behemoth of providing free information on the internet. It has a vast range of articles, from those that are written about events that happened hundreds of years ago, to breaking news stories that we witnessed in [3]. This allows further research into how this network of information can map the real world, and provide clues into solving some of the world's most pressing issues. One of those issues is health care. In this paper, we looked at three papers focusing on how Wikipedia's network of health articles, linked with various algorithms such as Google matrix, PageRank, and reduced Google matrix algorithms. We learnt that it can actually portray the real world and find hidden links between various cancers that researchers have never thought of. If [4], [5], and [6] can use the Wikipedia network to determine where various diseases such as cancers occur, it can help the entire health industry focus their resources on specific areas of the world to treat these types of cancers. For example, in [5], we saw that by using Wikipedia's network and PageRank, the authors were able to accurately map Wikipedia's network to WHO's disease study precisely. As algorithms become better at analyzing networks, we can gain better insights and information about some of the world's critical unsolved problems. This of course is a two way street. The algorithms can be amazing, but if the data is not accurate, wrong conclusions can be found. However, all three of these articles demonstrated that Wikipedia's network is correct, it is trustworthy, otherwise the conclusions that they came to would not have been possible. This is similar to our work on trying to determine communities within Wikispeedia's data Wikipedia's links [12]. After analyzing table 4 and understanding the trends, most of the categories in each community made sense. Articles that tend to be linked together remained in the same groups and certain categories were prominent in areas that you would expect them to be. Despite the large number of nodes in the first three communities, there were significant patterns that can be seen with just a single number indicating the number of nodes in each category. This further illustrates the importance of community recognition algorithms, such as the one that we used as it allows us to understand our network better. By seeing which articles interact with each other, we can devise better approaches in studying the data and make better decisions for what the final results should be. If that is building a better Wikipedia, or just optimizing it for the end users, data similar to this can help us build better applications for our users. Again, the content from Wikipedia was spot on, allowing us to find strong communities within the data, despite the unfortunate reality of many articles not given a category. This leads us back to the discussion on quality. In [3] and [11], the authors looked at how collaboration becomes to be in Wikipedia. From the experienced to novice, breaking news to old history, every article brings hundreds to thousands of editors, all trying to make sure that the content that is published is accurate and their best work. Yes, there are vandals, and they will always be, however we can implement better policies and procedures to stop them before they negatively impact the site. So is Wikipedia untrustworthy? If anything, the content on there is as accurate as it can be, it is just our perception that needs to be changed.

6. Conclusion

To conclude, we have discussed 10 related papers, created a network made up of Wikipedia internal links, analyzed the core statistics of that network, found the nodes that top the network with respective to their degree, out-degree, and PageRank, ran a community detection algorithm and analyzed the results, and finally discussed the impact of this paper. For some of the limitations, we would have liked to explore more statistics that made up the network such as betweenness, closeness, strongly and weakly connected components, and more. The other limitations are regarding the community assessment section. We would have liked to add categories to all unclassified articles as this could have displayed more statistics about each community. We felt that this definitely hindered the overall objective of this paper. For future work, instead of just comparing the categories of the articles, we would have liked to spend more time finding the exact link in each article and understand why that link matches with another article that might not share resemblances. This could allow us to build a better Wikipedia, one that can be optimized further for each unique user.

REFERENCES

[1] Anon. 2022. Size of wikipedia. (April 2022). Retrieved April 18, 2022 from https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[2] B.Thomas Adler, Luca de Alfaro, Ian Pye, and Vishwanath Raman. 2008. Measuring author contributions to the Wikipedia. *Proceedings of the 4th International Symposium on Wikis - WikiSym '08* (September 2008), 1–10. DOI:http://dx.doi.org/10.1145/1822258.1822279

[3] Brian Keegan, Darren Gergle, and Noshir Contractor. 2012. Do editors or articles drive collaboration? *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12* (February 2012), 427–436. DOI:http://dx.doi.org/10.1145/2145204.2145271

[4] Guillaume Rollin, José Lages, and Dima L. Shepelyansky. 2018. World influence of infectious diseases from Wikipedia network analysis. *IEEE Access* 7 (February 2018), 26073–26087. DOI:http://dx.doi.org/10.1101 /424465

[5] Guillaume Rollin, José Lages, and Dima L. Shepelyansky. 2019. Wikipedia network analysis of cancer interactions and world influence. *PLOS ONE* 14, 9 (September 2019). DOI:http://dx.doi.org/10.1371/journal.pone.0222508

[6] Guillaume Rollin, José Lages, Tatiana S. Serebriyskaya, and Dima L. Shepelyansky. 2019. Interactions of pharmaceutical companies with world countries, cancers and rare diseases from Wikipedia network analysis. *PLOS ONE* 14, 12 (December 2019). DOI:http://dx.doi.org/10.1371/journal.pone. 0225500

[7] Jakob Voss. 2005. Measuring Wikipedia. *International Conference of the International Society for Scientometrics and Informetrics* (April 2005).

[8] Joshua E. Blumenstock. 2008. Size matters. *Proceeding of the 17th international conference on World Wide Web - WWW '08* (April 2008), 1095–1096. DOI:http://dx.doi.org/10.1145/1367497.1367673

[9] Jun Liu and Sudha Ram. 2018. Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering* 115 (February 2018), 80–93. DOI:http://dx.doi.org/10.1016/j.datak.2018.02.004

[10] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. Measuring article quality in Wikipedia. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07* (November 2007), 243–252. DOI:http://dx.doi.org /10.1145/1321440.1321476

[11] Myshkin Ingawale, Amitava Dutta, Rahul Roy, and Priya Seetharaman. 2013. Network analysis of user generated content quality in Wikipedia. *Online Information Review* 37, 4 (2013), 602–619. DOI:http://dx.doi.org/10.1108/ oir-03-2011-0182

[12] Robert West and Jure Leskovec:Human Wayfinding in Information Networks. 21st International World Wide Web Conference (WWW), 2012.

[13] Robert West, Joelle Pineau, and Doina Precup: Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. 21st International Joint Conference on Artificial Intelligence (IJCAI), 2009.